Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Plant Biology

# We aren't good at picking candidate genes, and it's slowing us down

Ivan Baxter

In order to gain a molecular understanding of the genetic basis for plant traits, we need to be able to identify the underlying gene and the causal allele for genetic loci. This process usually involves a step where a researcher selects likely candidate genes from a list. The process of picking candidate genes is inherently prone to distortion due to human bias, and this is slowing down our research enterprise.

**Address**
Donald Danforth Plant Science Center, United States

Corresponding author: Baxter, Ivan (IBaxter@danforthcenter.org)

**Current Opinion in Plant Biology** 2020, **54**:57–60

This review comes from a themed issue on **Genome studies and molecular genetics**

Edited by **Joshua Cuperus** and **Christine Queitsch**

https://doi.org/10.1016/j.pbi.2020.01.006



As our ability to collect phenotype data has dramatically increased, we have not tracked the trajectory of data availability with tools that can translate that data into knowledge relevant to causal genes. This problem is not limited to genetics — research productivity, measured as a function of the number of researchers, has declined across many fields [1]. One obstacle has been that while our ability to collect massive amounts of data has improved, it is still cost-prohibitive to organize that data to make the connections that unambiguously assign gene function.

Advances in phenotyping have allowed for measurements of traits on hundreds to a few thousand plants, and field-based phenotyping has enabled the measurement of some traits over thousands of plots. This capacity is well suited for quantitative genetics populations of a few hundred to a few thousand lines. Leveraging the genetic variation that exists in nature or breeding programs and coupled with our ability to use modern sequencing methods to identify thousands to millions of polymorphisms, these techniques have identified thousands of marker trait associations [2,3].

We cannot overcome the limitations of linkage genetics so easily. Each marker is linked to multiple genes, and if the causal polymorphism is in a regulatory region, the gene it is controlling could be even farther away from the marker. So while we have been able to create large numbers of marker-to-phenotype associations, the next step of unambiguously assigning the functional gene is still quite difficult.

We might — *might* — be able to technology our way out of this. Maybe we do not need to know the genes. Instead, we can use genomic selection or prediction combined with AI to enable computers to determine the way to improve our crops [4–6]. Or, if we still want to get to the individual genes that underlie the trait of interest, we can combine cheaper and more powerful phenotyping with completely sequenced large populations. Mutagenesis creates a relatively small number of polymorphisms, and sequencing allows us to know all of them. If we find two independent lines with the same phenotype, whatever gene is mutated in both lines is the one causing the phenotype [7,8]. But we have a long way to go, and this approach will only find genes in which a single gene disruption causes the phenotype. *There are many uses where single-gene disruptions would be incredibly valuable*, but many genes will not have a phenotype when disrupted due to redundancy and epistasis. Knocking out a gene is a blunt tool. Polymorphisms that result in expression or biochemical alterations and their resulting phenotypes will enable an improved understanding of gene function.

As a result, the use of natural or breeding derived populations still has a great deal of value. To leverage that knowledge, we need to find ways to proceed from a marker-trait association to the underlying genes and alleles driving the phenotype (or trait). The current best practice is fine mapping through making targeted crosses and repeated rounds of genotyping and phenotyping. This process is hard enough for a few loci, particularly if phenotyping is done outside a controlled environment and the locus of interest has a strong GxE interaction, as most will [9]. This approach becomes completely untenable when scaling to thousands of associations. With a wealth of additional data sources available for many species, many researchers try to narrow the list of possible genes underlying loci by comparing that list to additional datasets to find genes with predicted mutations, tantalizing expression patterns, or other characteristics that may be related to the trait. If the researchers are correct in their hypotheses

about the potential meaning of these overlaps among datasets, this process can narrow down the list of candidates. However, this process rarely produces one and only one candidate, and it can lead to false positives if the assumptions used to select the datasets are wrong. So not only do researchers need to select the correct complementary datasets, but they also need to sift through the results to find the 'right' genes to follow up on. These two steps require human judgment, and I would submit that the well-known problems of human bias (especially, for example, confirmation bias) interferes with progress.

There is an idealized version of science in which scientists take an unbiased path in pursuit of answers. In reality, biases are abundant and significantly affect how our science proceeds. The most important biases, those that cause us to ignore and impede the progress of other researchers [10•,11], are not the subject of this piece. We are also, however, biased in how we choose directions and carry out research, and this limits our ability to identify the causal alleles of genes driving phenotypic variation in plants.

The clearest evidence for these biases in how we conduct research comes from the world of human research. Stoeger *et al.* [12•] looked comprehensively at which genes are studied by charting publications on human genes and comparing those with 430 physical, chemical, and biological features of the genes. Only 15 properties were required to predict the number of publications on a given gene, and those properties all affect the ability of genes to be studied by traditional methods (for example, 'number of tissues with detectable expression' or 'protein abundance in HeLa cells'). As a result, 49% of publications in 2015 focused on the 16% of genes that were known in 1991, almost a decade before the publication of the human genome. The authors identify several reasons that these biases may occur, including the availability of gene-specific resources for known genes and the belief that pursuing unknown targets carries a greater career risk (the authors find support for this fear in NIH data on authors who transition to PI positions). Fundamentally, when researchers have to choose which genes to spend more effort on, they tend to choose genes we already know about.

Is this situation better in plant research? I submit that it is probably worse, for several reasons. A primary one is that there is significantly more research conducted on human genes than on any plant species. Plant genomes are also less complete, and their annotations are of poorer quality, both from the level of predictions of gene structure [although efforts are under way to fix them [13]] and in the gene product information contained in GO terms (Box 1). As a result, much of the fascinating diversity of plant genes remains unexplored.

**Box 1 GO terms: our flawed — and best — resource for annotations**

Gene ontology (GO) terms represent our best attempt at converting human knowledge about genes into computationally readable formats. The structured formats of assigning biological processes, cellular components, and molecular functions allow for agreed-upon terms to be assigned to a gene along with an annotation of the evidence used to make the assignment. This improves the reliability and reproducibility of annotations, which are a key part of any human attempt to evaluate which genes are likely to account for their phenotype. Until recently, most genes in plant genomes did not have high-quality (by evidence code) GO annotations assigned and poor-quality annotations moved between genomes by automated annotation algorithms. Building and combining methods for annotation from multiple sources have enabled improved annotations for more genes [19]. Additionally, as with anything that humans do, assignment of GO terms involves bias, and humans are not good at assigning the correct GO terms. Fortunately, improvements in language processing may soon allow algorithms to assign GO and other terms relevant to phenotype after reading the text of manuscripts [20]. Efforts are also under way to leverage -omics-level data to improve GO annotations [21]. Despite these limitations, GO terms are currently the best, and in some cases the only tool available to test hypotheses about enrichment of gene sets and make inferences about potential causal genes relevant to phenotypes. Intensive efforts toward improving GO terms across all species are warranted.

There are also biases in how we choose to conduct our research, many of which are grounded in our responses to our own biology. For example, it is possible to time-shift plants through the use of programming growth chambers to environment regimes out of synch with local diurnal patterns, or to time-shift humans to work unusual schedules. However, these solutions either require extra effort and expense or researchers to sacrifice work-life balance. As a result, most experiments are done on plants during the day, which leaves our knowledge of the biology of the night somewhat, ahem, in the dark. Grinevich *et al.* [14•] illustrated this problem by demonstrating that the short-term transcriptional response to heat stress is dramatically different depending on what time of day that stress is imposed. Additionally, when heat-responsive genes were divided into groups based on nonstressed expression levels, multiple GO terms were enriched among genes with higher expression in the morning and genes with similar morning/evening expression levels. The heat-responsive gene sets where evening expression was higher (787 genes) had no significant GO enrichments. We would not be able to understand genes if we do not study them.

The above factors are surely not a comprehensive list of the way biases still affect our research directions. Unfortunately, these biases, like many ideas in research, become baked in, and it may take entirely new generations of researchers to circumvent them [15]. To overcome them, we need both to acknowledge that these biases exist and to find ways to actively combat them. We have many data-rich resources to utilize; the key is to find

ways to combine them to nominate candidate lists small enough to be manageable and to create resources that allow for rapid testing of those predictions. These data can come from a variety of sources — phylogenetic, epigenetic, gene function annotation, phenotypic descriptions, and so on, — but obvious candidates include proteomic and transcriptional datasets, which have gene-level data to identify specific candidates [16,17]. The downside is that it is important to select transcriptional/proteomic datasets that are in relevant tissues/populations/stresses, and biases in selecting these can lead researchers down the wrong path. There may also be ways to incorporate machine learning approaches to identify the genes [18], but it will be important to ensure that the training datasets are not overly reliant on lists created from biased approaches (for example, only looking at expression datasets collected during daylight). For resources, we need to bolster what is available in many species so that specific predictions in a species can be tested in that species. This is not to say that we should abandon working in models. Models will always have better resources, and identifying and characterizing genes in models is probably the best way to identify candidates in species with fewer resources. Ultimately, the goal is methods that identify groups of candidate genes small enough that all of them can be tested, removing the need for human selection and the biases that come with it.

## Conflict of interest statement

Nothing declared.

## Acknowledgements

Research in my lab is sponsored by the US National Science Foundation (IOS-638507) and the US Department of Energy (DE-SC0018277). I would like to thank Carolyn Lawrence-Dill, Blake Meyers, and the anonymous reviewers for constructive comments.

## References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest

1. Bloom Nicholas, Jones Charles I, Van Reenen John, Webb Michael: *"Are Ideas Getting Harder to Find?" Working Paper Series*. National Bureau of Economic Research; 2017 http://dx.doi.org/10.3386/w23782.

2. Tanger Paul, Klassen Stephen, Mojica Julius P, Lovell John T, Moyers Brook T, Baraoidan Marietta, Naredo Maria Elizabeth B *et al.*: **Field-based high throughput phenotyping rapidly identifies genomic regions controlling yield components in rice**. *Sci Rep* 2017, **7**:42839.

3. Prado Santiago Alvarez, Cabrera-Bosquet Llorenç, Grau Antonin, Coupel-Ledru Aude, Millet Emilie J, Welcker Claude, Tardieu François: **Phenomics allows identification of genomic regions affecting maize stomatal conductance with conditional effects of water deficit and evaporative demand**. *Plant Cell Environ* 2018, **41**:314-326.

4. Krause Margaret R, González-Pérez Lorena, Crossa José, Pérez-Rodríguez Paulino, Montesinos-López Osval, Singh Ravi P, Dreisigacker Susanne *et al.*: **Hyperspectral reflectance-derived relationship matrices for genomic prediction of grain yield in wheat**. *G3: Genes Genomes Genetics* 2019, **9**:1231-1247.

5. dos Santos Jhonathan PR, Fernandes Samuel B, Lozano Roberto, Brown Patrick J, Buckler Edward S, Garcia Antonio AF, Gore Michael A: **Novel Bayesian networks for genomic prediction of developmental traits in biomass sorghum**. *bioRxiv* 2019 http://dx.doi.org/10.1101/677179.

6. Baseggio Matheus, Murray Matthew, Magallanes-Lundback Maria, Kaczmar Nicholas, Chamness James, Buckler Edward S, Smith Margaret E, DellaPenna Dean, Tracy William F, Gore Michael A: **Genome-wide association and genomic prediction models of tocochromanols in fresh sweet corn kernels**. *Plant Genome* 2019, **12** http://dx.doi.org/10.3835/plantgenome2018.06.0038.

7. Addo-Quaye Charles, Tuinstra Mitch, Carraro Nicola, Weil Clifford, Dilkes Brian P: **Whole genome sequence accuracy is improved by replication in a population of mutagenized sorghum**. *G3: Genes Genomes Genetics* 2018, **8**:1079-1094 http://dx.doi.org/10.1534/g3.117.300301.

8. Li Guotian, Jain Rashmi, Chern Mawsheng, Pham Nikki T, Martin Joel A, Wei Tong, Schackwitz Wendy S *et al.*: **The sequences of 1504 mutants in the model rice variety kitaake facilitate rapid functional genomic studies**. *Plant Cell* 2017, **29**:1218-1231.

9. Li Zhi, Tirado Sara B, Kadam Dnyaneshwar C, Coffey Lisa, Miller Nathan D, Spalding Edgar P, Lorenz Aaron J *et al.*: **Characterizing allele-by-environment interactions using maize introgression lines**. *bioRxiv* 2019 http://dx.doi.org/10.1101/738070.

10. • Committee on the Impacts of Sexual Harrassment in Academic Science, Engineering, and Medicine: **Committee on Women in Science, Engineering, and Medicine, Policy and Global Affairs, and National Academies of Sciences, Engineering, and Medicine**. *Sexual Harassment of Women: Climate, Culture, and Consequences in Academic Sciences, Engineering, and Medicine*. Washington (DC): National Academies Press (US); 2018.
A comprehensive report of a pervasive problem.

11. Vettese Troy: *"Sexism in the Academy." n+1*. . May 2 2019 2019 *https://nplusonemag.com/issue-34/essays/sexism-in-the-academy/*.

12. • Stoeger Thomas, Gerlach Martin, Morimoto Richard I, Nunes Amaral Luís A: **Large-scale investigation of the reasons why potentially important genes are ignored**. *PLoS Biol* 2018, **16**: e2006643.
A detailed look at why some genes get more attention.

13. Tello-Ruiz Marcela K, Marco Cristina F, Hsu Fei-Man, Khangura Rajdeep S, Qiao Pengfei, Sapkota Sirjan, Stitzer Michelle C *et al.*: **Double triage to identify poorly annotated genes in maize: the missing link in community curation**. *bioRxiv* 2019 http://dx.doi.org/10.1101/654848.

14. • Grinevich Dmitry O, Desai Jigar S, Stroup Kevin P, Duan Jiaqi, Slabaugh Erin, Doherty Colleen J: **Novel transcriptional responses to heat revealed by turning up the heat at night**. *Plant Mol Biol* 2019, **101**:1-19.
The authors demonstrate how different sets of genes respond to stress depending on the time of day the stress is applied.

15. Azoulay Pierre, Fons-Rosen Christian, Graff Zivin Joshua S: *"Does Science Advance One Funeral at a Time?" Working Paper Series*. National Bureau of Economic Research; 2015 http://dx.doi.org/10.3386/w21788.

16. Diepenbrock Christine H, Kandianis Catherine B, Lipka Alexander E, Magallanes-Lundback Maria, Vaillancourt Brieanne, Góngora-Castillo Elsa, Wallace Jason G *et al.*: **Novel loci underlie natural variation in vitamin e levels in maize grain**. *Plant Cell* 2017, **29**:2374-2392.

17. Schaefer Robert J, Michno Jean-Michel, Jeffers Joseph, Hoekenga Owen, Dilkes Brian, Baxter Ivan, Myers Chad L: **Integrating coexpression networks with GWAS to prioritize causal genes in maize**. *Plant Cell* 2018, **30**:2922-2942.

18. Lin Fan, Fan Jue, Rhee Seung Y: **QTG-finder: a machine-learning based algorithm to prioritize causal genes of quantitative trait loci**. *bioRxiv* 2019 http://dx.doi.org/10.1101/484204.

19. Wimalanathan Kokulapalan, Friedberg Iddo, Andorf Carson M, Lawrence-Dill Carolyn J: **Maize GO annotation—methods, evaluation, and review (maize-GAMER)**. *Plant Direct* 2018, **2**:e00052.

20. Braun Ian R, Lawrence-Dill Carolyn J: **Automated methods enable direct computation on phenotypic descriptions for novel candidate gene prediction**. *bioRxiv* 2019 http://dx.doi.org/10.1101/689976.

21. Kramer Michael, Dutkowski Janusz, Yu Michael, Bafna Vineet, Ideker Trey: **Inferring gene ontologies from pairwise similarity data**. *Bioinformatics* 2014, **30**:i34-42.